University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Licensure Testing: Purposes, Procedures, and Practices

Buros-Nebraska Series on Measurement and Testing

1995

11. Equating

Judy A. Shea
University of Pennsylvania, SHEAJA@MAIL.MED.UPENN.EDU

John J. Norcini
American Board of Internal Medicine

Follow this and additional works at: https://digitalcommons.unl.edu/buroslicensure

 Part of the Adult and Continuing Education and Teaching Commons, Educational Assessment, Evaluation, and Research Commons, and the Other Education Commons

Shea, Judy A. and Norcini, John J., "11. Equating" (1995). *Licensure Testing: Purposes, Procedures, and Practices*. 16.
https://digitalcommons.unl.edu/buroslicensure/16

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Licensure Testing: Purposes, Procedures, and Practices by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

From: LICENSURE TESTING: PURPOSES, PROCEDURES, AND PRACTICES, ed. James C. Impara (Lincoln, NE: Buros, 1995). Copyright © 1995, 2012 Buros Center for Testing.

11

EQUATING

Judy A. Shea

University of Pennsylvania

John J. Norcini

American Board of Internal Medicine

INTRODUCTION

Testing programs nearly always need examinations that measure the same thing, but are composed of different questions (i.e., alternate forms of the same test). When different questions are used, however, there is no assurance that scores on the forms are equivalent; different sets of items might be easier or harder and, therefore, produce higher or lower scores. Equating is used to overcome this problem. Simply stated, it is the design and statistical procedure that permits scores on one form of a test to be comparable to scores on an alternate form.

A hypothetical example will help explain why equating is needed. Suppose Fred takes a certifying examination for aspiring baseball umpires. The examination has 100 questions sampled from the domain of questions about baseball rules and regulations. Fred gets 50 questions right and receives a score of 50. Ethel also takes an examination about baseball rules and regulations, but her test is composed of 100 different items. Ethel gets 70 questions right. Does Ethel know more about baseball than Fred? Or, might it be that Fred's test was much more difficult than Ethel's test, and contrary to appearances, Fred knows more about baseball than Ethel? The answers to these questions lie in equating, the process of ensuring that scores from multiple forms of the same test are comparable.

Equating is a technical topic and it generally requires a considerable background in statistics. The goal of this chapter is to provide a helpful and readable introduction to the issues and concepts, while highlighting useful references that

will provide technical details. The chapter begins with some general background and then presents common equating designs and an overview of methods and statistical techniques. For the most often used design, the common-item design, discussion will be expanded and examples will be provided. This will be followed by a consideration of factors that affect the precision of equating and an outline of some basic research questions. Finally, examples of currently available software will be inventoried.

BACKGROUND

At the outset it should be noted that the term "equating" implies that scores from different forms of a test will be rendered interchangeable. In fact, few data sets ever meet all of the strict assumptions that lead to interchangeable or equated scores. A more technically correct term would be scaled or comparable scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). In keeping with this notion, an attempt has been made to use the terms "scaled" or "comparable" scores throughout the chapter.

Reasons for Multiple Forms

There are at least three reasons to have multiple forms of a test. The first is security. Many testing programs administer high-stakes examinations in which performance has an important impact upon the examinee and the public: conferring a license or certificate to practice a profession, permitting admittance to a college or other training program, or granting credit for an educational experience. For a test score to have validity in any of these circumstances, it is crucial that it reflect the uncontaminated knowledge and ability of the examinees. Therefore, security is a concern and it is often desirable to give different forms to examinees seated beside each other, those who take the examination on different days, or those who take the examination on more than one occasion (Petersen, Kolen, & Hoover, 1989).

A second and related reason for different test forms is the current movement to open testing. Many programs find it necessary or desirable to release test items to the public (Holland & Rubin, 1982a). When this occurs, it is not possible to use the released items on future forms of a test without providing examinees an unfair advantage.

A third reason for different forms is that test content, and therefore test questions, by necessity changes gradually over time. Knowledge in virtually all occupations and professions evolves and it is crucial for the test to reflect the current state of practice. For example, it is obvious that today's medical licensure and certification examinations should include questions on HIV and AIDS, whereas these topics were not relevant several years ago. Even when the knowledge does not so obviously change, the context within which test items are presented is at risk of becoming dated. One could imagine a clinical scenario in medicine where descriptions of a patient's condition should be rewritten to include current drugs; in law one might want to include references to timely cases and rulings, especially

if they lead to different interpretations of the law. It sometimes happens also that the correct answer to previously used questions simply changes. When this occurs it is necessary to rewrite or replace the item. [As will be discussed later, equating assumes that the test scores are based on parallel forms of the test. Thus, if the changes in content are too severe, it is not appropriate to equate.]

Reasons to Equate

Given that different forms of an examination are necessary, it is important to ensure that the scores on one form of the test have the same meaning as the scores on another form. This issue of equivalence is important in most educational endeavors, but it is crucial in licensure and certification. Differences in pass/fail decisions across forms will undermine the meaning of a license or certificate. For example, through the 1970s, medicine was very popular and, according to some observers, it attracted the best and the brightest students. As medicine in general became less attractive in the 1980s, the quality of students entering, and therefore finishing, medical school may have declined. Without a method for ensuring the equivalence of pass/fail decisions on the licensing examination over time, students who passed in 1975 might have been more able than those who passed in 1990. This could have created "vintages" of licensed physicians. The license would not reflect the same standard over time and to know what it meant, it would be necessary to determine when a physician was granted the license. Consider as well, how unfair that would have been to the physicians seeking licensure. Some of those who were not good enough in 1975 would be by 1990 and vice versa.

Thus, the primary reasons for requiring equivalence are maintenance of the meaning of licenses/certificates and fairness to examinees. As stated in Lord (1977) (and later paraphrased in *Standards for Educational and Psychological Testing* [AERA, APA, & NCME, 1985]), "Transformed scores y* and raw scores x can be called 'equated' if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y" (p. 128). If this condition is met, it is then possible to make comparisons that are of interest to testing programs: among performances of different examinees who took alternative test forms, and among items and overall test scores that are given to various groups. A caveat is that in most cases, particularly those common to licensure and certification settings, equating is meant to adjust for unintended differences in form difficulty. As such, the real burden of creating parallel forms falls to test development. Thus, it is imperative that test developers and psychometricians collaborate to achieve the goal of producing interchangeable scores (Brennan & Kolen, 1987).

Conditions for Equating

In its simplest form, the process of equating has two components: selection of a data collection design and transformation of scores using a specific set of statistical techniques and methodologies. As will be discussed later in the chapter, there are several sound alternatives to choose among for both of these components. However, it is important to be acquainted with the four basic requirements or conditions for equating: (1) the different forms of the test should measure the same

attribute, (2) the resulting conversion should be independent of the data used in deriving it, (3) scores on the tests, after equating, should be interchangeable in use, and (4) the equating should be symmetric (Angoff, 1971/1984). Cook and Eignor (1991), Dorans (1990), and Petersen et al. (1989) provide very clear and extensive discussions of these requirements.

COMMON EQUATING DESIGNS

The first step in equating two forms of an examination is selection of a design. This involves two joint considerations: specifying which forms will be given on which occasions, and specifying which examinees will take which examination forms. Optimally, equating and the issues related to it will be a prospectively considered and integrated part of any testing program that must compare the performances of examinees and examinations over time. When equating details are not prospectively built into a testing program, it may sometimes be possible to change standard operating procedures to create a strong equating design. More often than not, however, the design that is actually used follows from the administrative procedures of the testing program already in place before the topic of equating becomes relevant (e.g., periodic administration to different groups of examinees, simultaneous administrations of several different test forms). Fortunately, adherence to already existing procedures is not a problem because several suitable equating designs exist.

Specification of a Design

Designs for equating vary along a continuum from straightforward to complex. Four basic designs serve as the building blocks of nearly all other commonly used strategies: (a) a single-group design—one group of examinees takes two (or more) forms of a test, (b) an independent groups and examination design—each examinee group takes a different form of the exam, (c) a counterbalanced design—each examinee group takes both (or all) forms of the exam, and (d) a common-item design—each examinee group takes a different form of the examination plus an anchor test composed of the same items. Each of these designs will be further explained below. In addition, more complex variations on these basic designs will be briefly presented. [Other authors conceptualize designs in somewhat different ways and they also use different terminology. See, for example, Petersen et al., 1989; Crocker & Algina, 1986]. Nevertheless, there is general consensus on which are the most basic designs.

Single group design. The simplest of the designs, though least practical by itself, is to give both (or all) forms of a test to a single group of examinees. The design could be portrayed as the following:

Group A
Form X +
Form Y

With this design, observed differences between test scores on the forms are due to differences in difficulty between the forms. In practice, this design is rarely used because it is difficult to convince examinees to take more than one form of an exam

and it is expensive to carry out as well. Even if examinees can be persuaded, scores on the second form may be contaminated by factors such as fatigue or practice. [There are ways to control for such unwanted effects; see the discussion below regarding counterbalanced designs.] Most importantly for licensure and certification settings, this design does not capture what actually happens in practice. That is, interest is most often in comparing scores for groups of examinees who take forms on different occasions or who take different forms, rather than looking at examination performance for two forms given at the same time.

Independent-groups design. A much more common situation is the one in which Examinee Group A takes Test X and Examinee Group B takes Test Y. For example, a licensing board might give an examination (Test X) in the fall of one year to one group of examinees (Group A) who just completed the required training for a profession. The next year a similar examination (Test Y) would be given to the new group of examinees (Group B) who recently completed their required training. The alternate forms are designed to be as similar as possible. In order to compare the performances of the two cohorts of examinees, psychometricians at the licensing board wish to transform the scores of one group (e.g., Group B) so that they are on the same scale as the other group (Group A). Schematically, the design would look like this:

Group A	<—	*Group B*
Test X		Test Y

This design would also apply when alternate forms are assigned to various examinees who take the examination simultaneously. For example, the design pertains when forms are assigned to examinees so that those sitting beside each other receive different tests.

When choosing an equating design, it is important to realize that no one direction of score transformation is inherently better than another. For example, with simultaneous administration of several forms, it is just as good to transform Test X scores so that they are on the Test Y score scale, as to transform Test Y scores so that they can be reported on the Test X score scale. However, in most licensure and certification settings, administrations occur over time. Thus, it makes most sense to report the more current scores on past scales; there is rarely a compelling reason to go back and change the scale on which earlier scores were reported.

Counterbalanced-groups design. The counterbalanced groups design is slightly more elaborate than the independent groups design. Both groups of examinees take both (all) forms of an examination. The presentation of forms would be counterbalanced (half of both examinee groups would receive Test X followed by Test Y and the other half would receive Test Y followed by Test X) to avoid factors such as practice and fatigue (Skaggs & Lissitz, 1986b). Schematically, the design would look like this:

Group A	—>	*Group B*
Test X +		Test Y +
Test Y		Test X

A design such as this is appealing because the comparability of forms can be directly evaluated; they are taken by the same group of examinees. For the same reason, examinees in various groups can be compared. However, like the single group design, this is rarely used in practice for obvious reasons: It is seldom practical to give examinees more than one complete test form because of limitations on examinees' and examiners' time and resources.

Common-item design. In contrast to designs that rely solely on the total test, an alternative is to adjust scores for examinees based on their performance on a set of common items that is administered to both groups. For example, Group A would take Exam X and Common Item Set U; Group B would take Exam Y and also take Common Item Set U. The schematic of this basic equating design could be more precisely specified as follows:

Group A	—>	*Group B*
Test X +		Test Y +
Common Set U		Common Set U

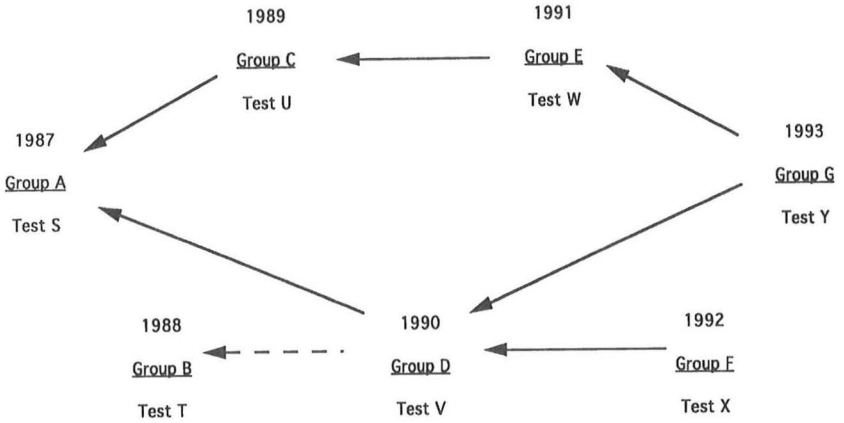
The common-item test, also called an anchor test, can be either external or internal to the focal test. Items that comprise the external anchor are usually not included in the examinees' reported test scores (Kolen, 1988). They are often presented as a separate section of the test, perhaps as a final test booklet. In contrast, with an internal anchor the common items are dispersed throughout the examination and are typically included as scored items that count towards the reported test score. The flexibility of the common-item design makes it useful in many different settings.

More complex designs. As mentioned in the introduction to this section, equating designs can be quite complex and often involve more than two examinations and two groups of examinees. Let us assume that a testing program that has been in existence for many years decides to begin equating examination scores. They have one administration per year and only one form of the examination is created for each administration Both of these procedures need to remain in place. In addition, it will be necessary to adhere to the longstanding policy that the same items never appear in two consecutive examinations. What might an equating plan look like for this organization?

For convenience, let us say that the base year will be 1987; examination scores in future years will be transformed to be on the same scale as this initial administration. In 1987 we will give Test S to Group A. In 1989, Group C takes Test U, which needs to be rescaled to Test S. Two years later, in 1991, Group E is administered Test W which is equated to Test S, through Test U. This pathway is shown in the top of the diagram.

Recall that items cannot be reused in consecutive years. The 1988 Test T will, of course, be given to Group B but it cannot be linked to the base year. However, in 1990 the Group D test takers can take Test V, which has items in common with both Test S and the 1988 form (which is being ignored in this diagram). Two years later, in 1992, Group F is administered Test X, which is equated to Test S, through Test V. This pathway for the even-numbered years is shown in the bottom of the

diagram. Finally, the 1993 examination will be "double-linked" to previous forms through the items it has in common with 1990 and 1991. A design such as this may be depicted as follows:



Designs such as these are referred to as chained or braided designs. One problem in the implementation of equating over time is that errors can accumulate. Such problems can be overcome by interlacing the groups/examinations at prespecified intervals. For example, in the diagram above, the 1993 form was linked to both 1990 and 1991. The reason for doing this is to insure that separate "strains" of the examination do not develop, such as an "even year" strain and an "odd year" strain. Note also, that the 1988 form and the 1992 form were not used in the current chain. However, both would be brought into future equatings via shared items with 1994, which might also share with 1991. Drawing schematics can help visualize how checks can be built into the system, as well as define what is practical for any particular organization. Literature evaluating these complex chaining or braiding designs is very useful for highlighting issues and problems that can occur over time and threaten the integrity of the equating (Petersen, Cook, & Stocking, 1983), as well as bringing out problems that cannot be detected in short-term designs and evaluations.

For the design above, one could add a common-item test to each administration. Moreover, the common-item set could change over time. That is, the common-item set used to link Test X and Test Y need not be the same as the common-item set used to link Test Y and Test Z. The implication of this is that the content of the common-item link is allowed to change over time to better reflect the goals of the testing program and to maintain security of the examination forms.

Another design that deserves mention is a preequating design. Originally discussed by Educational Testing Service (Holland & Rubin, 1982b), preequating refers to inclusion of different groups of items in multiple examination forms. The preequated items are not included in the examinees' test scores but the necessary data are collected to allow calculation of equating transformations. The preequating can be done in either an item (Kolen & Harris, 1990) or section format (Holland

& Thayer, 1985). Preequated items are then subsequently assembled into a form(s) and administered at a later date. Preequating permits rapid scoring when the time between administration of the forms and deadlines for reporting results is short. Also, implementation of a preequating design builds in some protection against administering a seriously flawed exam. A possible design for one administration might look like the following:

Group A	*Group B*	*Group C*
Test X +	Test X +	Test X +
PE Form A	PE Form B	PE Form C

Eventually, preequated (PE) Forms A, B, and C would be put together to form Test Y. The transformations would be calculated prior to administration and when Test Y is administered, it could immediately be reported on the Test X scale. The preequated forms would not contribute to examinees' test scores at the initial administration. Naturally, however, one would want the PE forms to look like other parts of the test so that examinees would apply equal effort. [The same holds true for any section of experimental or pretested questions that is not included in examinees' scores.]

The number of other designs that could be developed is large, as are the statistical techniques for performing the equatings. Fox example, preliminary methods have been developed for multidimensional equating (Hirsch, 1989) and equating with confirmatory factor analysis (Rock, 1982). At this point, these technically demanding procedures have not gained widespread use.

In sum, the specific design and direction of equating that one chooses will be closely intertwined with the more general structure, policies, and procedures of the testing program. The most important points in the discussion of design are: (a) design simply refers to how data are collected from various examinees and, (b) there are four simple designs that serve as building blocks for more complex structures. The remainder of the chapter will utilize the Independent-Groups and Common-Item Designs, the most typical equating situations (Cook & Eignor, 1991).

Selection of Examinees

In the process of defining an equating design it is necessary to specify the sample of examinees who will take the forms on which the equating transformations will be based. The most important consideration in designating equating subsamples is whether they are random or nonrandom selections of examinees (some authors refer to equivalent and nonequivalent groups, see Dorans [1990] or Petersen et al. [1989]). Several designs call for the selection of random samples of examinees to receive various test forms (see Angoff, 1971/1984) because it is reasonable to assume that they are of equivalent ability. However, in practice it is usually not feasible to do this and, more often than not, the structure of the testing environment and practical considerations dictate that the samples will be nonrandom.

A second issue, independent of the random-nonrandom decision, is specifying exactly which examinees will be included in the equating subsamples. Examinees involved in equating need not necessarily be all those who take a particular form at a particular administration (Harris, 1987). It is best to select fairly large groups

of examinees, who exhibit some variability in performance but whose skills and training are relatively homogeneous. That is, even though groups cannot be precisely equivalent, efforts are made to create groups that are as comparable as possible.

Emphasizing homogeneity may mean omitting some test takers. For example, many testing programs allow examinees to take multiple administrations of the exam, either because they are trying to better earlier performance (e.g., MCAT scores, GREs), or because they failed to meet established pass-fail or cutoff points. In these instances, it is better to limit the equating transformations to first-time takers of the examination, because they tend to have known training and educational experiences. Similarly, one might not want to include examinees who are admitted to the examination following unusual training or educational experiences, or those who elect to take the examinations at various times of the year. Several investigators have found sizable performance differences between examinees taking spring and fall administrations of an examination (Cook & Petersen, 1987; Petersen et al., 1983; Schmitt, Cook, Dorans, & Eignor, 1990). Seasonal shifts have also been reported for a medical licensing examination (Nungester, Dillon, Swanson, Orr, & Powell, 1991). Whatever the final decisions regarding selection of examinees, the samples used in equating should be well justified and explained to all interested parties (AERA, APA, & NCME, 1985).

A third consideration in selecting or describing samples of examinees relates to deciding whether they differ only slightly in ability, or whether they differ considerably. The former is referred to as horizontal equating, and is applicable in most testing programs where the abilities of the examinees remain fairly constant from one administration to another (e.g., examinees sitting for licensure and certification examinations). The latter is referred to as vertical equating and is quite common in programs such as educational achievement and aptitude testing programs where there is a desire to compare scores for examinees at different grades or training levels. In horizontal equating, the tests are designed to be similar and differ for only unintended reasons. In vertical equating, the tests are intentionally designed to differ in difficulty (Cook & Eignor, 1983).

Technically, many of the procedures for horizontal and vertical equating are the same. The practical difference is that the accuracy and precision of equating are typically much greater in the case of horizontal equating (Skaggs & Lissitz, 1986b). However, even in large, ongoing testing programs in which horizontal equating should suffice, there may be subtle but consistent changes in the examinees' abilities over several administrations. For example, examinees sitting for certification in internal medicine showed consistent declines in performance over a period of several years (Norcini, Maihoff, Day, & Benson, 1989). Admittedly, the distinction between horizontal and vertical equating designs is not always clear. Nevertheless, asking the question focuses attention on expected examinees' abilities and helps to elucidate the equating procedure and anticipated equating results.

In sum, selection of designs and examinees was considered separately because it is important that the issues relevant to each be considered. In practice, many discussions of equating describe various designs by jointly specifying how the

samples of examinees and the selection of items or forms occurred. One of the most widely known typologies was provided by Angoff (1971/1984). Among the designs he describes are Design I: Random groups—one test administered to each group; Design II: Random groups—both tests administered to each group, counterbalanced; Design III: Random groups—one test administered to each group, common-equating test administered to each group; and Design IV: Nonrandom groups—one test to each group, common-equating test administered to both groups. Familiarity with this work provides a very thorough background and is helpful when reading current literature.

EQUATING METHODS AND PROCEDURES

Having chosen a design and the examinees, it is necessary to select a method for transforming the scores from the various forms to be on the same scale. Specific transformation or equating procedures fall within two psychometric theories: conventional (traditional) test theory and item response theory. Within traditional test theory there are several equating methods. The most common and well studied are the equipercentile method and linear equating methods. In contrast, methods falling under the rubric of item response theory (IRT) have only been widely discussed for the past 10 to 15 years, but they are proliferating rapidly. At this point, the IRT models that have received the most attention in the published literature are the one-parameter (Rasch) and three-parameter models, based on logistic estimation procedures (Baker, 1985; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Wright & Stone, 1979). However, marginal maximum likelihood estimation procedures (Bock & Aitkin, 1981) are becoming quite popular.

The focus of this section of the chapter will be on general assumptions and equating methods that can be associated with estimation procedures from either conventional or item response theory. The interested reader is referred to the references listed above for more extensive discussions. In addition, this discussion assumes that test scores are sums of dichotomously scored items (right/wrong). Methods for other types of data are just becoming widely available (Baker, 1992; Thissen, 1991). As such, existing equating methodologies are, for the most part, not yet useful for clinical data or data derived from item formats that produce other than 0/1 responses.

Traditional Test Theory

Equipercentile equating. Equipercentile equating is a method of transforming scores so that, when the equating is complete, two scores are said to be comparable if they have the same percentile or rank within their respective examinee group. This method makes no statistical assumptions about the tests to be equated. However, the result is that the distributions underlying each form are identical in all moments (i.e., they have the same distribution). The procedure stretches or compresses the two distributions so that this outcome is achieved.

Equipercentile equating is typically done by computer, though it is relatively easily done by hand. The general procedure has several steps and application to an

independent-groups design is sketched below. For a thorough and detailed example, the reader is referred to Angoff (1971/1984). Procedures are slightly more complicated for common-item designs; see Angoff (1971/1984), Dorans (1990), and Thorndike (1982) for descriptions of alternative procedures.

1. A distribution of test scores is developed in a tabular format, and percentile ranks or relative cumulative frequency distributions are prepared. This is done separately for each form of the examination taken by a different group of examinees.
2. The cumulative distributions for each form are plotted on a graph and each graph is smoothed. Smoothing, as the term suggests, is the process of transforming the sometimes jagged curve that is produced by plotting actual distributions to a "smooth" curve. In the past, smoothing was done by hand. It can also be done analytically, for example by the rolling weighted average method (Angoff, 1971/1984), or any of several very sophisticated methods detailed by Fairbank (1987), Hanson (1991), and Kolen (1991).
3. Once the distributions are plotted and smoothed, a table is made showing the raw scores from each form that correspond to several different percentiles. For example, the table would show what score from Form X and what score from Form Y correspond to a percentile rank of 85. This is repeated for many (usually about 30) other percentile ranks. Rather than selecting every possible percentile, the investigator may select many smaller increments in the part(s) of the distribution where he or she is most interested in precision. Also, numerous closely spaced points will have to be taken at both ends of the distributions where scores are rare.
4. A second graph is made showing the relationship between pairs of scores entered into the table above. If necessary, this graph is also smoothed.
5. From the final graph, a table is made showing the appropriate conversions between the two test score distributions. For example, it might show that a score of 5 on Exam X is equivalent to a score of 4 on Exam Y.

The major advantage of the equipercentile technique is that it is quite suitable for describing curvilinear relationships between scores on different tests. But, a fairly significant disadvantage that causes many investigators to choose other models is that the process of smoothing is quite subjective. Moreover, this method forces distributions of two scores to be the same, even when there may be legitimate reason for having very different distributions (i.e., the purpose of the examination changes and it becomes more or less difficult). As Cook and Petersen (1987) discuss, this method is entirely data dependent. If other observed distributions of test scores were equated, a different conversion table would emerge. This is likely to be particularly true at the tails of the distribution where there are few data points. Clearly, large samples are needed for precise equating. On the other hand, with large samples that sometimes occur in licensure and certification programs, scores

will be observed over the entire range including the area that contains the cutting score. Overall, the equipercentile method has been widely used and continues to be the preferred method for some testing programs (e.g., American College Testing). In some sense, it remains the standard against which other methods are compared.

Linear equating. The second common equating procedure is linear. The general formula that applies is a linear transformation of the form Y = AX + B, where A and B are parameters that use standard score terms to express the ideas of equating [$(x - m_x)/s_x = (y - m_y)/s_y$], X refers to scores on Test X, and Y refers to scores on Test Y (Petersen et al., 1983). This general linear formula is applicable in many different designs, among them being Angoff Designs I through IV described above (Angoff, 1971/1984). However, the designs differ in the way in which the transformation constants, A and B, are calculated.

The computational formulas appropriate for a common-item design with nonrandom groups (Designs IVa in Angoff parlance) are shown below to illustrate how straightforward the linear equating process is. This example was selected because it represents the most common scenario in licensure and certification testing: Different forms of an examination are administered on different testing occasions. The derivation of the formulas, attributed to Tucker (Gulliksen, 1950), is presented in Angoff (1971/1984). The goal is to calculate the coefficients that fulfill the equation Y = AX + B where $A = s_{y_t} / s_{x_t}$ and $B = M_{y_t} - AM_{x_t}$ The four equations to be solved are:

$$M_{x_t} = M_{x_a} + b_{xu_a}(M_{u_t} - M_{u_a})$$

$$M_{y_t} = M_{y_b} + b_{yu_a}(M_{u_t} - M_{u_b})$$

$$s^2_{x_t} = s^2_{x_a} + b^2_{xu_a}(s^2_{u_t} - s^2_{u_a})$$

$$s^2_{y_t} = s^2_{y_b} + b^2_{yu_b}(s^2_{u_t} - s^2_{u_b})$$

Where:

M_{u_t} = the observed mean of Groups A and B on the Common Set U

M_{u_a} = the observed mean of Group A on Common Set U

M_{u_b} = the observed mean of Group B on Common Set U

M_{x_a} = the observed mean of Group A on Exam X

M_{y_b} = the observed mean of Group B on Exam Y

b_{xu_a} = the regression coefficient from regressing Exam X scores for Group A on Common Set U scores

b_{yu_b} = the regression coefficient from regressing Exam Y scores for Group B on Common Set U scores

$s^2_{u_t}$ = the observed variance of Groups A and B on Common Set U scores

$s^2_{u_a}$ = the observed variance of Group A on Common Set U scores

$s^2_{u_b}$ = the observed variance of Group B on Common Set U scores

$s^2_{x_a}$ = the observed variance of Group A on Exam X scores

$s^2_{y_b}$ = the observed variance of Group B on Exam Y scores

The list of all of the components for the equations is long, but calculation of the appropriate terms and the ultimate transformation of scores can quite easily be done with standard software packages such as SPSS (Norusis, 1992), SAS (SAS Institute, Inc., 1989), SYSTAT (Wilkinson, 1992), and BMDP (Dixon, 1990). The two examples below are based on applications to typical testing situations and they illustrate how easy the computations can be.

Returning to the example of hypothetical scores on the baseball rules and regulations test, it is possible to illustrate what is involved in equating, and in fact, why equating is necessary. Recall that Fred (as a part of Group A) received a score of 50 on the Form X 100-item baseball test. Ethel (as a part of Group B) received a score of 70 on the Form Y 100-item baseball test. The question to answer is how these scores compare to one another. Ultimately, a direct comparison can be made after the scores for Group A Test X are transformed to be on the same scale as the Group B Test Y scores. For the moment, we will forget about the performance of individuals and focus on group statistics.

Scenario #1—tests of different difficulty. Assume that in addition to their respective 100-item Forms, Groups A and B also took the same 30-item set of common items, referred to as Test U. Performance on the form-specific items (often called "unique" items in the literature) and the common items might look as follows:

	X Mean	X *SD*	U Mean	U *SD*
Group A	60	7	15	3
	Y Mean	Y *SD*	U Mean	U *SD*
Group B	65	8	11	5

Before the equating is done, some observations can be made from these data that foreshadow the results after equating. Notice that Group B did not score nearly so well on the common items as Group A, even though their scores on the form-specific items (Test Y) were somewhat higher. This suggests that Group A Test X scores will in all likelihood be "raised" when they are transformed to the Group B Test Y scale.

Proceeding with the equating will clarify the relationship. Other computations (not shown here) indicate that the combined performance of Groups A and B on the Common Item Set U has a mean of 13 and a standard deviation of 4. The result of regressing Group A Test X scores on Group A Common Set U scores is .90. The result of regressing Group B Test Y scores on Group B Common Set U scores is .80. These are all of the data that are needed to complete the equating transformation. Into the formulas given previously we substitute the following:

$$M_{x_t} = M_{x_a} + b_{xu_a}(M_{u_t} - M_{u_a})$$
$$= 60 + .90(13 - 15) = 58.20$$
$$M_{y_t} = M_{y_b} + b_{yu_a}(M_{u_t} - M_{u_b})$$
$$= 65 + .80(13 - 11) = 66.60$$

$$s^2_{x_t} = s^2_{x_a} + b^2_{xu_a}(s^2_{u_t} - s^2_{u_a})$$
$$= 7^2 + .90^2(4^2 - 3^2) = 54.67$$
$$s^2_{y_t} = s^2_{y_b} + b^2_{yu_b}(s^2_{u_t} - s^2_{u_b})$$
$$= 8^2 + .80^2(4^2 - 5^2) = 58.24$$

Further substitution results in the equating coefficients:

$$A = s_{y_t} / s_{x_t}$$
$$= 58.24^{1/2}/54.67^{1/2} = 1.03$$
$$B = M_{y_t} - AM_{x_t}$$
$$= 66.6 - 1.03(58.2) = 6.65$$

Thus, Y = AX + B becomes Y = 1.03X + 6.65.

Glancing at the formula tells us that roughly 6 or 7 points need to be added to all Group A Test X scores before they can be compared to Group B Test Y scores. More precisely, Fred's score of 50 is transformed to 58.15 (Y = 1.03(50) + 6.65). Thus, his score is lower than Ethel's score but not as much as it originally appeared. This result should be reassuring to all Group A test takers. Test developers might want to ask why Form X is more difficult than Form Y.

Scenario #2—Examinee groups with different ability. This time let us assume that Group B had the better performance on the common items. The scores are:

	X Mean	X *SD*	U Mean	U *SD*
Group A	60	7	11	5
	Y Mean	Y *SD*	U Mean	U *SD*
Group B	65	8	15	3

What will happen to Fred's score in this case? Intuitively, one might guess that the Group A scores will be lowered. Not only do they score lower on a similar test, they do much worse (about a standard deviation worse) on the common items.

As before, the regression coefficient regressing Group A Test X scores on Group A Common Set U scores is .90. The regression coefficient regressing Group B Test Y scores on Group B Common Set U scores is .80. When these data are appropriately substituted into the equations the resulting linear equation is: Y = 1.28X - 15.70. That is, Fred's score of 50 becomes a 48.30. The intuitions were correct and all Exam X scores will be lowered. In this scenario, test developers and administrators would do well to ask why the apparent ability of the two groups was different. Did they set out anticipating group differences (i.e., vertical equating) or is some selection or training factor creating the differences? Perhaps the licensure or certification examination is becoming more or less attractive to certain groups of examinees.

Creating scenarios such as the two presented is a very helpful learning tool. Those involved with equating may find it useful to create other scenarios that represent their own testing situation. For example, if the equating groups have equal mean performances, but their variances are very different, there will be an

obvious and predictable impact on equating (i.e., the score distribution will expand or contract depending on which is chosen as the base form). Similarly, the equating transformation is influenced by the degree of correlation between the anchor test and the whole test forms (Budescu, 1985). As a postscript it should be noted that the examples provided above were hypothetical, and numbers were chosen for ease of calculation. In actual testing situations, the process may be a bit less straightforward.

A clear advantage of linear equating methods is their ease of implementation. Also, linear methods do not have subjective components such as the equipercentile method does with smoothing. On the other hand, they are fairly simplistic and assume that a simple linear equation is sufficient to describe the relationship between score distributions.

Common-item equating also depends on making a number of statistical assumptions. They are spelled out in Braun and Holland (1982), Kolen and Brennan (1987), and Petersen et al. (1989). The two assumptions that receive the most attention and are the most readily testable are: (a) linearity of the regression of the whole test form score on the anchor test score, and (b) homogeneity of the residual variation about the regression (Braun & Holland, 1982). Other requirements depend on the specific mathematical transformation being utilized. For example, Thorndike (1982) says that equating must involve equally precise (i.e., reliable) tests, and that both (all) tests should have the same correlation with a third measure. In contrast, Angoff (1971/1984) presents formulas for tests of unequal reliability. It is advisable that users of test equating procedures become familiar with the specific assumptions of the techniques under consideration (or in use). Petersen et al. (1983) present a very helpful table comparing the widely used Tucker and Levine methods. It is important to repeatedly perform checks to assess how well the test data continue to meet the assumptions of the model.

Item Response Theory

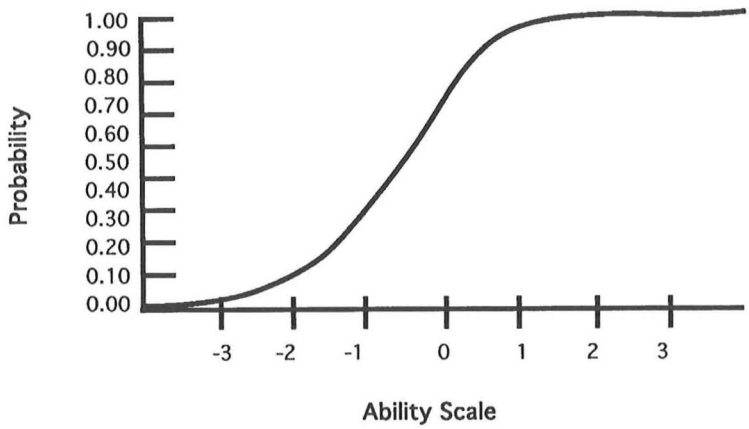
Item response theory (sometimes called latent trait theory) has been increasingly studied over the past 10 to 20 years (Wright & Stone, 1979; Lord, 1980; Hambleton & Swaminathan, 1985). The goal of the theory is to model performance on a trait using observed test scores. There are numerous item response theory models, developed from competing mathematical frameworks (Birnbaum, 1968; Bock & Aitkin, 1981; Swaminathan & Gifford, 1982, 1983). The most basic IRT model, often referred to as one-parameter model, says that performance on a particular item is a function of the examinee's ability and the difficulty of the item. More complex models add item discrimination to the prediction model (two-parameter model) and a chance or guessing factor (three-parameter model). Before a discussion of equating within IRT can occur, it is helpful to (a) review some basic concepts from item response theory and (b) contrast IRT with traditional test theory. More extensive discussion of the models is beyond the scope of this chapter. For more detail the reader is referred to Hambleton and Swaminathan (1985), Hambleton et al. (1991), Lord (1980), and Baker (1985).

General concepts. The usual outputs of IRT calibrations are sets of item parameters and estimated person (i.e., examinee) abilities. Item parameters are

conceptually analogous to item statistics in that they describe features of an item: the b-parameter refers to the difficulty of an item, the a-parameter refers to discrimination, and the c-parameter is a pseudo-guessing factor. However, it is important to note that IRT parameters are not numerically or statistically equivalent to traditional item statistics. Similarly, person (examinee) abilities, expressed as thetas with standard errors, quantify how well each person performed, though they do not appear as, nor are they equivalent to, raw scores.

Estimation of item and person parameters is generally an iterative process, occurring is successive stages until an acceptable amount of precision is reached (the termination values are determined by various software programs and can be adjusted by the user). In the end, one obtains a matrix of item parameters (with standard errors attached to each parameter) and a vector of estimated person abilities (with standard errors for each estimate). Because item and person parameters are jointly estimated, they are placed on the same [arbitrary] scale within a calibration run. In one popular program (BILOG, Mislevy & Bock, 1989), the estimates of person ability have a range of -3 to +3 and are centered on 0 with a standard deviation of 1. Item difficulties are centered above or below this mean, depending on if the items are generally difficult for the average test taker (above) or easy (below) for the average test taker. Item discrimination varies between 0 and infinity, though the upper range is usually set at around +2. The pseudo-guessing factor varies between approximately 0 and the reciprocal of the total number of item choices (e.g., .20 for an item with five answer choices).

Information from the estimated item parameters and ability estimates is portrayed in item characteristic curves (ICC), the building blocks for all IRT models. This focus on individual items is a significant departure from conventional test theory where the focus is on total test scores. An ICC is a plot describing how the characteristics of an item interact with a person's ability. Stated another way, it is a graph showing the probability of a correct response to a particular item over the entire ability range. Usually it is an S-shaped curve with the examinee ability scale along the abscissa and the probability of a correct response on the ordinate. A sample is shown below.



Curves located to the right of the midpoint of the ability distribution represent difficult items whereas curves located to the left of the midpoint represent easy items. Steep curves indicate highly discriminating items. Lower asymptotes above 0 suggest that guessing is influencing estimates for the lowest ability examinees.

ICCs are summed over all items to create test characteristics curves that describe the function of all items over all test takers. Finally, with IRT one is able to calculate information functions. This is a measure of the precision of estimation for each item over the entire ability range. Information functions can be summed over all items to create test information functions (TIF). TIFs identify at what point(s) in the ability distribution of examinees, information is maximized for a set of items. Roughly, information is inversely related to the standard error of estimate for person ability. If most test items match the ability of the examinees *and* the items are highly discriminating, the test characteristic curve will be a steeply peaked curve, the peak representing the point on the ability scale where the test provides the most information. If the majority of the items do not match the average ability of examinees, the curve will be very flat, suggesting the test provides minimal information along the ability distribution.

Comparison to traditional test theory. Traditional test theory is based on what Hambleton and Swaminathan (1985) describe as a set of weak assumptions. Because the assumptions are weak, the theory is applicable in most typical testing situations. On the other hand, tests based on traditional test theory have some shortcomings: (a) the item statistics (*p*-values and *r*-biserials) apply only to the specific group who took the examination on which the scores are calculated; (b) comparisons of scores are limited to situations where examinees take parallel examinations; and (c) it is presumed that scores are equally precise over the entire range of ability.

In contrast to traditional test theory, item response theory is based on a set of very strong assumptions. First, it is assumed that the test data are unidimensional, meaning that they measure only one trait or ability (multidimensional models have been developed but they are not widely used at this time) (Hambleton & Swaminathan, 1985). Second, the data must exhibit local independence (Lord, 1980). Simply stated, this is the requirement that for examinees of the same ability, responses to particular items are uncorrelated. Third, it is assumed that the test is not speeded. The one-parameter model also requires that all items in an examination be equal in discrimination and that "guessing" by examinees does not influence responses to any items. Clearly, this is quite a stringent set of assumptions.

Additional reading in item response theory will show that many early studies focused on assessing data-model fit for particular data sets (Hambleton & Murray, 1983; Shea, Norcini, & Webster, 1988), comparing techniques for investigating fit of the models to the data (Hambleton & Rovinelli, 1986), or investigating how robust the models were to violations of the assumptions (Dorans & Kingston, 1985). As with methods resulting from conventional test theory, there is rarely a clear answer to the question of "how much misfit is too much?" However, sizable departures from unidimensionality and equal item-total discrimination are rela-

tively easy to spot. When violations do occur, the user should select a more complete model, or use conventional equating methodologies.

For all models, when the observed test data appropriately fit the model, item response theories theoretically offer several advantages over conventional test theory. The advantages that are most relevant to equating are that estimates of examinees' abilities are independent of the particular sets of items on which the ability estimates are based, and similarly, estimates of item parameters (i.e., difficulty and discrimination) are independent of the particular set of examinees on whom the item parameter estimates are based. For example, proponents of the theory would suggest that if all test items were divided into odd-even numbered subsets, or easy-hard subsets, the same ability estimates would be obtained for an examinee regardless of which subtest he/she took. Similarly, estimated item parameters are theoretically the same for subsamples of examinees such as highest and lowest ranked class members, or first-time test takers and repeaters (though they will have to be rescaled by a constant because scaling within a single IRT run is arbitrary).

Other advantages are also present with IRT, such as more accuracy in transformation at the extremes of the scale. Also, because IRT statistical manipulations are conducted at the item level, rather than the total test score level, IRT offers the possibility of item preequating (e.g., deriving equating transformation data before an operational form is actually administered) (Cook & Eignor, 1983; 1991). At this point it is appropriate to reiterate a previously stated caution. The advantages of IRT described above are achieved if, and only if, the model of interest fits the actual test data. In reality, this rarely occurs. Moreover, high quality calibration of item and person parameters requires larger sample sizes than linear equating methods, especially when the common joint maximum likelihood estimation procedures are used.

Equating procedures. For purposes of this discussion, assume the data to be equated adequately fit the model(s) of interest. How then, does one equate? As discussed by Cook and Eignor (1991) IRT equating is a three-step process: (a) select a design, (b) place parameter estimates from different samples on a common scale, and (c) equate test scores. The issue most relevant for equating becomes selecting the appropriate methodology for placing item parameters on the same scale.

In general, there are three methods for transforming item parameters generated from different samples of examinees to be on the same scale. The most straightforward is concurrent calibration. Data for multiple examination forms and examinees are simultaneously calibrated and scaled within one computer run, thus the item and ability estimates are automatically on the same scale (i.e., Steps 2 and 3 are completed simultaneously). This method would probably be the ideal but limitations on computer resources make this procedure impractical on occasion. Moreover, if items are calibrated following one test administration and performance is reported to examinees, it does not usually make much sense when the next administration occurs to recalibrate the items taken by the original sample.

The alternative equating methods use a common-item design. The first of these alternatives is called the fixed-b design. In this method, all items for one examination

form are calibrated (i.e., the *a*s, *b*s, and *c*s are estimated as are the person abilities). Then, the item parameter estimates for the common items, in particular item difficulties, are entered as fixed values into the subsequent run for the second form. All other (non-common) items (Step 2) and all ability estimates (Step 3) will be scaled around these preset values.

A second alternative is to employ a rescaling technique based on the relationships between item parameters estimated for common-item links. The simplest rescaling procedure, applicable only when the data meet the assumptions of the Rasch model, calculates the mean item difficulties for the two sets of common items, estimated independently (Wright & Stone, 1979). The difference in the means is computed and this value is added to the difficulty estimates (Step 2) *and* ability estimates (Step 3) for the test form to be transformed (Baker, 1985; Wright & Stone, 1979).

Another common-item alternative, appropriate regardless of the IRT model, is referred to as the mean and sigma method (Hambleton & Swaminathan, 1985). Ability and item estimates are transformed using the equation $y = Ax + B$, where $A = s_y/s_x$ and $B = y - Ax$. The *A*s and *B*s are used to transform estimated item difficulties ($b_i^* = Ab_i + B$), item discriminations ($a_i^* = a_i/A$), and ability estimates $\Theta_a^* = A\Theta_a + B$.

Variations on the mean and sigma method include the robust mean and sigma methods proposed by Linn, Levine, Hastings, and Wardrop (1981) and Stocking and Lord (1983). These variations take into account the accuracy of estimation and give less weight to outliers among the common items. Similarly, a second method proposed by Stocking and Lord (1983), referred to as the characteristic curve method, improves on the basic linear procedure by making use of the discrimination parameter and the entire ability distribution in addition to the difficulty parameter in calculating the transformation coefficients. Thus, theoretically it could be expected to result in a more exact transformation.

It is beyond the scope of this chapter to report and evaluate these alternative transformation techniques (see McKinley [1988] for a comparison of several methods). However, there is an abundant literature that makes comparisons among the various IRT procedures as well as between IRT and conventional equating methods (e.g., Baker & Al-Karni, 1991; Skaggs & Lissitz, 1986a).

Finally, a note should be made about Step 3. The procedures for placing parameter estimates on a common scale are also used to transform ability estimates. If it is tenable to report rescaled ability estimates on a theta scale (typically ranging from -3 to +3), then the equating procedure is complete. In most cases, however, it is necessary to translate the theta estimates for both forms to a scale that makes more sense to examinees (i.e., corresponding estimated true scores). For example, examinees and other interested parties may be accustomed to seeing scores reported on a scale with a mean of 500 and a standard deviation of 100. If it is important to maintain this scale, the procedures and an example for doing this final transformation are provided in Cook and Eignor (1991).

In sum, there are many potential benefits of item response theory that support test equating. The need to meet the strict assumptions of these models has already

been mentioned and should not be dismissed. More practically, the largest disadvantage is the unfamiliarity of both testing professionals and consumers with the theory. Equally important is the lack of research clearly supporting the utility of a particular IRT methodology. Although each theory has its supporters, as does each method of parameter transformation (usually linked directly to a particular software program), it is not at all clear when the potential benefits accrued from using IRT outweigh the uncertainties. For the time being, conventional methods are a better choice and there is unlikely to be an appreciable loss of precision in licensure and certification examinations due to their use. Cook and Eignor (1983) offer a very clear discussion of the basic issues.

Comparison of Equating Procedures

During the 1980s and early 1990s there have been numerous studies comparing the outcomes of various equating techniques in horizontal and vertical equating settings. A complete review cannot be provided here; the reader is referred to Petersen et al. (1983) and Skaggs and Lissitz (1986b) as examples of excellent reviews and methodologies.

Overall, several authors have concluded that when the tests to be equated are similar in content and difficulty, and the design describes a horizontal equating situation, IRT methods are neither consistently better nor worse than conventional methods. Both conventional and IRT methods work well, particularly the three-parameter logistic model (Lord, 1980; Marco, Petersen, & Stewart, 1983; Petersen et al., 1983). When the tests do differ in content and length, or the anchor test differs from the remainder of the test(s), some authors have found that methods based on the three-parameter item response model perform better (e.g., Petersen et al., 1983) whereas others support use of conventional methods (e.g., Skaggs & Lissitz, 1986a). In part, the differences among studies are due to how the tests were designed, whether the data were real or simulated, and the choice of criterion. Current research results do not consistently support, at least from a psychometric perspective, the superiority of any one method. In fact, as noted by Skaggs and Lissitz (1986b) "it is unreasonable to expect a single equating method to provide the best results for equating all types of tests" (p. 495).

Conclusions regarding vertical equating are more straightforward. Most, though not all, researchers have concluded that vertical equating is problematic for both conventional and IRT methods, particularly the one-parameter Rasch model (Harris & Hoover, 1987; Loyd & Hoover, 1980; Gustafsson, 1979). See Harris (1991) and Skaggs and Lissitz (1988) for exceptions.

How should a researcher then choose a procedure, given the breadth of research results? Theoretically, IRT has some appeal *if* the data meet the assumptions of the model(s). The assumptions must be tested thoroughly; they cannot be assumed to be met. Further, it is doubtful that typical data produced by certifying and licensure examinations would provide adequate fit with the one-parameter IRT model. IRT methods require expertise in actually using the techniques, as well as in explaining them to interested users and consumers. At this point, few licensure and certifying bodies have ready access to individuals with the

training to use IRT methods appropriately, although if an agency is just embarking on equating, it is probably as easy to learn IRT methods as conventional methods.

In summary, there are few differences among methods when examinations are parallel and examinees are of nearly equal ability. Conventional methods have the advantages of being easier to apply, understand, and explain to consumers. Consequently, without compelling reasons to the contrary, conventional methods are preferable. What should actually happen is that testing organizations should compare the two classes of methods to determine which fits their situation the best.

FACTORS AFFECTING THE PRECISION OF EQUATING

Numerous factors affect the precision of equating. Consistent results over many studies suggest general guidelines that might be followed in initiating and maintaining an equating program. Topics pertinent to a common-item design are listed below. Few authors study all facets simultaneously.

Anchor Test Length

A rule of thumb for many years has been that the common-items link should be roughly 20% the length of the total test or 20 items, whichever is longer (Angoff, 1971/1984). For conventional equating, lengths over 20 items seem not to have an advantage if the examinee groups are similar in ability (Klein & Kolen, 1985; Norcini, 1990). For IRT, some researchers have reported that much shorter anchor tests (as few as two or five well-chosen items) work well (Raju, Bode, Larsen, & Steinhaus, 1986; Vale, 1986). However, other researchers working within IRT suggest 15 to 20 items are more appropriate (Hills, Subhiyah, & Hirsch, 1988; Wingersky, Cook, & Eignor, 1986). Unless there is a persuasive need for a very short anchor, in light of the equivocal results regarding length, the 20% guideline still seems sensible.

Content Representation

One of the most widely cited studies with the anchor test design is by Klein and Jarjoura (1985). They investigated differences between content-representative anchors and longer, but nonrepresentative anchors; all anchors were matched to the total test in terms of difficulty. They included two different equating methods and results were evaluated with several different statistics. Overall, they found that content representation was very important for accurate equating results, especially when the groups of examinees were nonrandom. These results were supported by Petersen, Marco, and Stewart (1982) who concluded from their comparison of numerous linear equating models that even moderate differences in content between an anchor and the total test led to substantial error.

Difficulty of Anchor Test

Another characteristic of anchors that is often studied is difficulty. That is, researchers ask about the effects on equating when the anchor test is, and is not, similar in difficulty to the scored test. Petersen et al. (1982) found that differences in difficulty between an anchor and the total test were related to substantial error. Similarly, in a companion piece comparing error of equating for conventional and

IRT equating methods, they found that differences in difficulty between the anchor test and the form-specific items resulted in substantial error for the linear methods investigated, especially when the samples of examinees differed in ability (Marco et al., 1983). However, it might be noted that the differences did not affect error for the IRT-based methods, nor in situations when the examinee samples were random.

Ability of Examinee Groups

Studies looking at the results of vertical equating are not particularly encouraging. Though vertical equating will typically not be a problem for licensure and certification agencies where approximately equivalent groups take examinations over time, there is ample research to suggest that even when the differences in ability between the groups involved in the equatings are small, the impact upon equating may be sizable (Angoff & Cowell, 1986; Petersen et al., 1982). It should be noted, however, that some authors have found that all commonly used models are fairly robust to differences in examinee ability (Harris & Kolen, 1986).

Examinee Sample Size

Almost as often as researchers have asked how many common items are needed, they have also asked how many examinees are needed. In one article, a minimum sample size of 400 was recommended (Brennan & Kolen, 1987) for conventional equating techniques. However, another study found that errors of equating were not appreciably bigger with samples of 250 than of 500 (Norcini, 1990). Similar results using linear equating were found for samples of 200, 300, and 400, even when the samples were disparate in ability (Shea, Dawson-Saunders, & Norcini, 1992). More strikingly, a recent study that combined sample sizes of 25, 50, 100, and 200 with various smoothing techniques applied to the equipercentile method suggested that very small samples could be appropriate in some situations. These results are not definitive but they should be encouraging to examiners who consistently deal with small groups of test takers (Livingston, 1993).

In contrast to conventional methods, it is generally accepted that large samples are necessary for some item response theory software packages. Cook and Eignor (1991) suggest that as many as 2,000 examinees are needed for stable initial item calibration with joint maximum likelihood calibration. Smaller samples (i.e., a few hundred examinees) are sufficient for other IRT estimation procedures, such as marginal maximum likelihood and Bayesian (Drasgow, 1989; Harwell & Janosky, 1991; Stone, 1992).

In sum, several studies have concluded that equating works best when the characteristics of the common items represent those of the total test. Though few authors have studied variations in content, difficulty, length, and ability groups simultaneously, it is generally recommended that the common-item set should mirror the total test in content and statistical properties (Cook & Eignor, 1991). In essence, the higher the correlation between the anchor and the test, the more effective the equating (Thorndike, 1982). This is certainly the most conservative approach, especially when outcomes of equating have a significant and immediate impact on examinees' professional lives.

From the foregoing discussion, it is fair to conclude that many of the potentially troublesome issues surrounding equating can be averted by sound test construction processes. Potthoff (1982) presents many test construction ideas, and raises issues that deserve thoughtful consideration. Brennan and Kolen (1987) similarly list test development guidelines.

ISSUES THAT NEED MORE RESEARCH

Throughout this chapter, several topics have been mentioned that warrant additional attention. Many of the topics were outlined by Brennan and Kolen (1987). A partial list would include the following topics.

Scale Drift

Several investigators have shown that drift occurs over time with linked/chained equatings (Cook & Eignor, 1983; Petersen et al., 1983). More research is needed to identify (a) the conditions under which scale drift does and does not occur and (b) the effectiveness of methods to prevent it.

Security Breaches

Security breaches are always a threat to the validity of examination scores; they are particularly relevant to equating when they involve items in a common-item link. Most certifying examinations are administered under relatively secure conditions. Nevertheless, examination books do turn up missing from time to time, or test takers become acquainted with specific items. Simulations that consider issues such as the number of items affected and the length of time until discovery (e.g., several administrations) would prepare test agencies for possible future needs.

Changes to the Common-Item Link

Inevitably, changes will occur in a common-item link. Perhaps it will be discovered that an item was miskeyed, or perhaps new discoveries in a particular field will require that the answer to an item changes. When this occurs, decisions need to be made about alterations to the common-item link and the impact that such alterations have on examinee scores. Dorans (1986) provides a detailed and thorough account of the impact of several possible decisions, depending on the characteristics of the item.

Location Effects for Anchor Items

Many authors have discussed the effect of location or context of items upon examinee performance (e.g., Cook & Petersen, 1987; Harris, 1991; Kingston & Dorans, 1984; Kolen & Harris, 1990). Most of the studies have not focused on internal common-items links, though Thorndike (1982) did note that anchor items should be presented to examinees taking different forms at the same points so that practice and fatigue could be avoided. Because the performance on anchor items is especially important in determining examinees' scores, the impact of location should be further investigated.

Rounding

The numerous texts and empirical papers on the topic of equating provide an abundance of formulas and examples. However, there appears to be little uniformity regarding how many decimal places are used throughout the statistical manipulations, and there is no mention at what stages rounding occurs. The implicit consensus is that it is best to work with maximum precision throughout the equating process, but this is not explicitly stated (for exceptions see Potthoff, 1982 and Brennan & Kolen, 1987). Hand calculations using the scenarios presented earlier show that level of precision can make a difference to examinees, particularly those who score near the cutting score.

Equating Based on Standard-Setting Judgments

To this point, the discussions of equating have assumed that the goal is to transform scores on a test form so that they are comparable to scores on an alternate examination form. In a licensure or certification situation, however, actual test scores are sometimes less important than pass-fail decisions. Nevertheless, the scores of all examinees are transformed as usual and the cutting score or pass-fail point is among the scores that are altered. The rescaled cutting score is then used to make the pass/fail decisions. This ensures that the same licensure or certification decisions are being made regardless of which form of a test is taken.

For some kinds of licensure and certification situations, however, score equating may not work very well. For example, when the number of examinees is small or the pass-fail point is located far from the mean, score equating does not work well (Brennan & Kolen, 1987). Conventional equating might also not be optimal when nontraditional testing formats are used (e.g., essays, performance tests), or testing time is limited so that long anchor tests are impractical.

Several recent studies by Norcini and colleagues (Norcini, 1990; Norcini & Shea, 1992; Norcini, Shea, & Grosso, 1991; Norcini, Shea, & Lipner, 1994) have sought to address this issue by applying a common-item design and a linear statistical technique to the data gathered when experts set standards. In other words, rather than inputting data from examinees' scores (mean, standard deviations, etc.) into the formulas listed above for the common-item design, the data that are used in the calculations are generated via application of a standard-setting technique to each item in an examination. Specifically, for many licensure and certification examinations, the pass/fail point is chosen using a variation on Angoff's standard-setting method (Angoff, 1971/1984). As part of this process, a group of experts meets and each makes judgments about the proportion of borderline examinees who would respond correctly to each item. The result of this procedure is that each judge has "scores" on the whole test and the anchor test. Statistics summarizing these scores over all judges can readily be put into the equating formulas. Cutting score equivalents produced by this method can be compared to the results obtained by traditional score equating and to a criterion.

The series of studies concluded the following:

1. Rescaling based on experts' judgments (approximately 8 to 10 judges per group) was more accurate than equating based on examinee samples of 100, 250, and 500, and performed about the same as equatings based on samples of 1,000 and 2,000 examinees.
2. Results were stable for 25 or more common items and 5 or more judges. The amount of error was approximately 1 item on a 100-item test. Increasing the number of common items or judges resulted in little improvement in precision.
3. Transformed Angoff values were stable when compared to original estimates and bias in the estimates was small. This implies that rescaled Angoff values could be included in an item bank, and equivalent pass-fail decisions would result regardless of which items were chosen for a particular form of the test.
4. Use of judges' estimates in a common-item design was robust to unusual, or at least mismatched, common-item links that were fabricated of items either high or low in difficulty, and high or low in discrimination.

In sum, this area of research is still in its infancy but the issue it raises, equating at the cutting score, has relevance for certifying and licensing organizations. Results of early studies are encouraging but need to be extended to other types of examinations and judges.

Criteria to Evaluate Equating

Criteria used to evaluate the outcomes of equating procedures vary from one investigation to another. In empirical studies, such evaluation is often done by equating a test to itself and looking at how much variation (drift) has occurred over the numerous equatings. Another strategy is to define a "gold standard" criterion based on logically and/or theoretically acceptable arguments. In either case, researchers are apt to evaluate how well equated scores meet their criteria by reporting mean differences, mean absolute differences, or root mean square errors. Although these results are often convincing and informative, they frequently do not address the needs in practice for evaluating equating results in ongoing testing programs. Skaggs and Lissitz (1986b) provide a very thoughtful discussion of the issue. Additionally, Kolen (1990) points out the wisdom of using a "no equating" condition as a criterion.

Standard Errors of Equating

Closely related to the topic of appropriate criteria is the issue of standard error. For several of the conventional, linear methods, standard errors of equating have been developed. See, for example, the discussion presented by Petersen et al. (1989). Similarly, Jarjoura and Kolen (1985) present a method for estimating standard errors of equipercentile equating. In their discussion they point out that use of an inappropriate method (i.e., a linear method for a curvilinear relationship) can be particularly troublesome at the extremes of the distribution, where cutting scores are often located.

Misfitting Items

If IRT methods are used with the common-item design, the psychometrician needs to expend considerable effort ensuring that the items in the link (as well as the total tests) meet the assumptions of the particular model under investigation. Though models may be robust with a few misfitting items (Cook, Eignor, & Wingersky, 1987), research has not defined the limits outside of which misfit will adversely affect the results.

Biased Items

Another aspect of equating that has only recently received attention is bias or differential item functioning (Candell & Drasgow, 1988; Linn et al., 1981). This is a question of whether the items perform differently than expected with certain subpopulations of examinees, for example, white and African American examinees, or men and women. If so, then such items should not routinely be included as common items (and should not even be in the test form). Cook et al. (1987) discuss the importance of making sure that none of the items in the anchor test are biased for any examinee subgroup.

Alternative Item Formats

There is a need to investigate optimal equating designs and statistical techniques for item types other than multiple-choice questions (MCQs). Certainly, MCQs remain representative of most testing programs. However, in many fields there is a desire to move away from MCQs towards new formats such as performance tests and simulations. Investigations are quickly needed to explore how equating can be performed with alternative formats such as standardized patients, essays, and portfolios that involve new issues such as multiple correct answers and longer testing times per "item," thus limiting the number of test items available.

Multidimensional Tests

Many examinations used in professional licensure and certification settings comprise multiple dimensions. Clearly this is a problem for the widely used IRT models. Exploratory IRT work has begun to address multidimensional equating (Hirsch, 1989), as well as determine how bias results from multidimensionality (Oshima & Miller, 1992), but the methods are not widely used. On the other hand, multidimensionality does not specifically pose a problem for equatings within conventional theory, if the tests to be equated are similarly multidimensional (Cook & Petersen, 1987).

Adaptive Testing

Throughout the testing field there is an increased emphasis on adaptive testing. (See Wainer, 1990, for a comprehensive overview.) Generally speaking, this is the procedure of administering different sets of items to each examinee, targeted to his or her ability level. Consequently, each examinee may take different subsets of items, and raw scores will not be directly comparable. A somewhat different issue,

but still presenting the same problem, is that of tailored testing. In tailored testing, examinees are allowed to select examination modules based on training and practice characteristics and interests.

It is not immediately clear how equating could be applied to adaptive testing using conventional equating techniques. An item bank in which all of the items have been placed on the same scale using IRT procedures presents one solution to these problems. However, issues of order and context effect could be potentially troublesome because the location of item presentation will undoubtedly be different for the calibration sample than for future examinees for whom the item is selected during an administration (Petersen et al., 1989).

Matching Examinee Samples on Ability

Angoff and Cowell (1986) have shown that even slight heterogeneity in the two equating groups can seriously impact on the equating transformations. A solution to this problem may lie in matching, that is, artificially improving the correspondence between the two examinee groups involved in the equating by matching on some examination score or external criterion (Dorans, 1990). A set of empirical studies in a special issue of *Applied Measurement in Education* (Wise, Plake, & Mitchell, 1990), using both real and simulated data (Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Schmitt et al., 1990), explored matching under several different conditions and with different methods. Though theoretically a sound idea, the results suggest that, at best, matching is risky (Kolen, 1990; Skaggs, 1990).

Other Issues

As one thinks about the test development and administration procedures for a specific testing program, in all likelihood issues that have not been discussed, and for which there is little research, will arise. For example, it many be necessary to give test forms in different languages. Or, examinees with special needs may require altered test administration procedures. A third example is the need to decide what to do when test administration procedures are nonstandard for some examinees (the electricity goes out, there is distracting noise around the testing site). At this point, research cannot suggest how to handle each of these unique events, except to reiterate that the purpose of equating is to construct test scores that are equivalent, thus insuring fairness to examinees. Adaptation of the best studied methods described in this chapter should provide helpful responses.

SOFTWARE OPTIONS

Performing the statistical transformations required for equating can be done by hand (or hand-calculator) if examinee samples are small and the less complex conventional linear procedures are used. However, for ongoing testing programs some type of software will almost always be needed.

With an examination scoring system already in use, and a desire to employ conventional linear methods, it is not too demanding to write programs for equating procedures using a standard statistical software package such as SPSS (Norusis,

1992), SAS (SAS Institute, Inc., 1989), SYSTAT (Wilkinson, 1992), or BMDP (Dixon, 1990), or, if the expertise is available, using a language such as Fortran or C. Alternatively, a relatively new program for the widely used common-item design is LEQUATE. The program can handle either internal or external anchors, and it implements two widely used linear procedures (Tucker and Levine) (Waldron, 1988). It runs on IBM/PC and compatible DOS-based PCs. Documentation and the program are available free of charge from William J. Waldron, Tampa Electric Company, P.O. Box 111, Tampa, FL 33601.

Within item response theory (IRT) there are many choices; the three most widely used to date are BICAL, LOGIST, and BILOG. Published reviews and comparisons of various software programs are often helpful in making a selection decision (e.g., Harwell & Janosky, 1991; Mislevy & Stocking, 1989; Stone, 1992).

BICAL was developed for the one-parameter (Rasch) item calibration and equating (Wright & Stone, 1979); as such it has relatively limited uses. It provides estimated item parameters (the b or difficulty parameter only) and person ability estimates. It uses maximum likelihood estimation procedures and is available for DOS-based PCs. In the past 20 years, the program has evolved from BICAL to newer versions called BIGSTEPS, MSCALE, and MSTEPS. BIGSTEPS is the currently recommended PC version; it can reportedly handle responses for 20,000 examinees and 3,000 items. Information and prices on the program can be obtained from MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637; (312) 702-1596 or (312) 288-5650 (phones); (312) 702-0248 (FAX).

LOGIST is a very comprehensive and flexible program, developed by Educational Testing Services. It uses maximum likelihood estimation procedures and the user can select the one-, two-, or three-parameter IRT models. A strength of this program is that it has been in use for many years so there is ample literature to read for educational and comparative purposes. It does require relatively large sample sizes for calibration. At this point it is only available for use on a mainframe but a personal computer version is forthcoming. Copies are available from Educational Testing Service, Rosedale Road, Princeton, NJ 08541.

BILOG has become a popular IRT alternative in recent years. It uses marginal maximum likelihood item parameter estimation procedures, and is capable of handling one-, two-, or three-parameter IRT models. Scale scores can be estimated with maximum likelihood, Bayes, or Bayes modal procedures. The program is available for DOS and OS-2 based systems. Recent versions for UNIX operating systems are also available and a Windows version is nearly ready for release. The user's manual is clear and helpful. Information regarding the software may be obtained from Scientific Software International, 1525 East 53rd Street—Suite 530, Chicago, IL 60615-4530, (800) 247-6113 (phone); (312) 684-4979 (FAX). SSI also offers several other IRT-based software programs appropriate for item formats other than dichotomously scored (right/wrong) items: BIMAIN, MULTILOG, PARSCALE, and TESTFACT.

With LOGIST and BILOG, equating can be achieved with concurrent calibration or the fixed *b*s method. However, if one is using a common-item design and does not wish to recalibrate at each administration, then another method will have

to be used to calculate the transformation constants and then rescale the estimated-item parameters and person abilities. One possibility that works reasonably well is to use a standard statistical software package, such as SPSS (Norusis, 1992), SAS (SAS Institute, Inc., 1989), SYSTAT (Wilkinson, 1992), or BMDP (Dixon, 1990), and do your own programming. An alternative is to get access to routines used by other investigators that were specifically designed for this purpose. Examples are EQUATE and EQUATE 2.0, programs written in FORTRAN for use on DOS-based PCs. EQUATE was developed for dichotomously scored items and uses the test characteristic curve method of equating. EQUATE-2 extends EQUATE capabilities to include graded or nominal scoring procedures. They were designed by Frank Baker and colleagues at the University of Wisconsin (Baker, Al-Karni, & Al-Dosary, 1991; Baker, 1993) and are available upon request from Frank Baker, Department of Educational Psychology, Educational Sciences Building, 1025 W. Johnson Street, University of Wisconsin, Madison, WI 53706.

Final examples that one might find useful are RASCAL and ASCAL, marketed by Assessment Systems Corporation. RASCAL computes item parameter estimates and person ability estimates within the one-parameter (Rasch) IRT model. ASCAL performs the same tasks for the two- and three-parameter models. RASCAL estimates are based on an unconditional maximum likelihood estimation procedure and ASCAL used Bayesian modal estimation. With RASCAL, the user can "fix" item difficulties to predetermined values. With ASCAL, the user can link (i.e., equate) items from different administrations onto a single scale during one run. Both programs run on DOS-based personal computers. They reportedly can handle up to 250 test items and several thousand examinees (30,000 for RASCAL and 15,000 for ASCAL).

A potential benefit of RASCAL and ASCAL for some users is that they can be integrated into a broader testing system called MicroCAT. MicroCAT is a relatively complete test-design and administration system. Within the multifunction system, it is possible to develop items (with graphics), print test forms, do item and test analysis, and create result report forms. If IRT is chosen for item analysis, items can be calibrated with RASCAL or ASCAL. Conventional item analysis (and thus, score equating) is also available. MicroCAT is available from Assessment System Corporation (2233 University Avenue, Suite 440, St. Paul, MN 55114). It is also available from SAGE Publications, Inc. (P.O. Box 5084, Thousand Oaks, CA 91359-9924). It might also be noted that the user can work with personnel at Assessment System Corporation to develop customized packages to meet one's particular needs.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Services. Originally published in R. L. Thorndike (Ed.),

(1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327-345.

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating procedures. *Journal of Educational Measurement, 28*, 147-162.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15*, 78.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-445.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic Press.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11*, 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*, 13-20.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

Cook, L. L., Eignor, D. R., & Wingersky, M. S. (1987, April). *The effect on IRT equating of using linking items with problematic item response functions.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal

circumstances. *Applied Psychological Measurement, 11*, 225-244.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Dixon, W. J. (Ed.). (1990). *BMDP statistical software*. Berkeley, CA: University of California Press.

Dorans, N. J. (1986). The impact of item deletion on equating conversions and reported score distributions. *Journal of Educational Measurement, 23*, 245-264.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*, 3-17.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249-262.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching with a fallible criterion. *Applied Measurement in Education, 3*, 37-52.

Fairbank, B. A., Jr. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11*, 245-262.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gustafsson, J. E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*, 153-158.

Hambleton, R. K., & Murray, L. M. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391-408.

Harris, D. J. (1991). A comparison of Angoff's design I and design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28*, 221-235.

Harris, D. J. (1987, May). *Effect of comparability of examinee groups*. ACT Research Report Series 87-5. Iowa City, IA: American College Testing.

Harris, D. J. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*, 247-256.

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151-159.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*, 35-43.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. *Journal of Educational Measurement, 25*, 221-231.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26*, 337-349.

Holland, P. W., & Rubin, D. B. (1982a). Introduction: Research on test equating sponsored by Educational Testing Service, 1978-1980. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 1-6). New York: Academic Press.

Holland, P. W., & Rubin, D. B. (Eds.) (1982b). *Test equating*. New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational Statistics, 10*, 109-120.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10*, 143-160.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197-206.

Klein, L. W., & Kolen, M. J. (1985, April). *Effect of number of common items in common-item equating with nonrandom groups*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*, 97-104.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*, 257-282.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*, 263-277.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*, 27-39.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*, 19-36.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23-39.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73-95.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1993). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147-177). New York: Academic Press.

McKinley, R. L. (1988). A comparison of six-methods for combining multiple IRT item parameter estimates. *Journal of Educational Measurement, 25*, 233-246.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Norcini, J. J. (1990). Equivalent pass/fail decisions. *Journal of Educational Measurement, 27*, 59-66.

Norcini, J. J., Maihoff, N. A., Day, S. C., & Benson, J. A., Jr. (1989). Trends in medical knowledge as assessed by the certifying examination in internal medicine. *Journal of the American Medical Association, 262*, 2402-2404.

Norcini, J. J., & Shea, J. A. (1992). Equivalent estimates of borderline group performance in standard setting. *Journal of Educational Measurement, 29*, 19-24.

Norcini, J., Shea, J., & Grosso, L. (1991). The effect of numbers of experts and common items on cutting score equivalents based on expert judgment. *Applied Psychological Measurement, 15*, 241-246.

Norcini, J. J., Shea, J. A., & Lipner, R. S. (1994). The effect of anchor item characteristics on equivalent cutting scores. *Applied Measurement in Education, 7*, 187-194.

Norusis, M. J., & SPSS, Inc. (1992). *SPSS for Windows. Base system user's guide.* Release 5.0. Chicago: SPSS, Inc.

Nungester, R. J., Dillon, G. F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard-setting plans for the NBME comprehensive part I and part II. *Academic Medicine, 66*, 429-433.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*, 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 221-262). New York: American Council on Education-Macmillan Publishing Company.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Academic Press.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 201-242). New York: Academic Press.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986, April). *Anchor-test size and horizontal equating with the Rasch and three-parameter models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Rock, D. A. (1982). Equating using the confirmatory factor analysis model. In P. W. Holland, & D. B. Rubin (Eds.), *Test equating* (pp. 247-257). New York: Academic Press.

SAS Institute, Inc. (1989). *SAS/STAT user's guide* (version 6, fourth ed.). Cary, NC: SAS Institute, Inc.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education*, *3*, 53-71.

Shea, J. A., Dawson-Saunders, B., & Norcini, J. J. (1992, April). *The effects of equating when examinee groups vary in size and ability.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Shea, J. A., Norcini, J. J., & Webster, G. D. (1988). An application of item response theory to certifying examinations in internal medicine. *Evaluation and the Health Professions*, *11*, 283-305.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education*, *3*, 105-113.

Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, *10*, 303-317.

Skaggs, G., & Lissitz, R. W. (1986b). IRT test equating: Relevant issues and a review of recent literature. *Review of Educational Research*, *56*, 495-529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, *12*, 69-82.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1-16.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*, 175-191.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter model. In D. Weiss (Ed.), *New horizons in testing* (pp. 13-30). New York: Academic Press.

Thissen, D. (1991). *MULTILOG user's guide, version 6.0.* Mooresville, IN: Scientific Software.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin, 1982.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.

Waldron, W. J. (1988). LEQUATE: Linear equating for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 12,* 323.

Wilkinson, L. (1992). *SYSTAT: Statistics (Version 5.2).* Evanston IL; SYSTAT, Inc..

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986, April). *Specifying the characteristics of linking items used for item response theory item calibration.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wise, S. L., Plake, B. S., & Mitchell, J. V., Jr. (Eds.) (1990). *Applied Measurement in Education, 3*(1).

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: Mesa Press.